

Construction and Iteration-Complexity of Primal Sequences in Alternating Minimization Algorithms

Quoc Tran-Dinh

Received: date / Accepted: date

Abstract We introduce a new weighted averaging scheme using “Fenchel-type” operators to recover primal solutions in the alternating minimization-type algorithm (AMA) for prototype constrained convex optimization. Our approach combines the classical AMA idea in [18] and Nesterov’s prox-function smoothing technique without requiring the strong convexity of the objective function. We develop a new non-accelerated primal-dual AMA method and estimate its primal convergence rate both on the objective residual and on the feasibility gap. Then, we incorporate Nesterov’s accelerated step into this algorithm and obtain a new accelerated primal-dual AMA variant endowed with a rigorous convergence rate guarantee. We show that the worst-case iteration-complexity in this algorithm is optimal (in the sense of first-order black-box models), without imposing the full strong convexity assumption on the objective.

Keywords Alternating minimization algorithm · smoothing technique · primal solution recovery · accelerated first-order method · constrained convex optimization

1 Introduction

This paper studies a new weighted-averaging strategy in alternating minimization-type algorithms (AMA) to recover a primal solution of the following constrained convex optimization problem:

$$f^* := \begin{cases} \min_{\mathbf{u}, \mathbf{v}} \{f(\mathbf{x}) := g(\mathbf{u}) + h(\mathbf{v})\} \\ \text{s.t. } \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}, \quad \mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}, \end{cases} \quad (1)$$

where $g : \mathbb{R}^{p_1} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^{p_2} \rightarrow \mathbb{R} \cup \{+\infty\}$ are both proper, closed and convex (not necessarily strongly convex), $(p_1 + p_2 = p, \mathbf{A} \in \mathbb{R}^{n \times p_1}, \mathbf{B} \in \mathbb{R}^{n \times p_2}, \mathbf{c} \in \mathbb{R}^n$, and $\mathcal{U} \subset \mathbb{R}^{p_1}$ and $\mathcal{V} \subset \mathbb{R}^{p_2}$ are two nonempty, closed and convex sets.

Problem (1) surprisingly covers a broad class of constrained convex programs, including composite convex minimization, general linear constrained convex optimization problems, and conic programs.

Quoc Tran-Dinh
Department of Statistics and Operations Research
The University of North Carolina at Chapel Hill, USA
E-mail: quocd@email.unc.edu

Primal-dual methods handle problem (1) together with its dual formulation, and generate a primal-dual sequence so that it converges to a primal and dual solution of (1). Research on primal-dual methods has been extensively studied in the literature for many decades, see, e.g., [4, 17, 19] and the references quoted therein. However, such methods have attracted a great attention in the past decade due to new applications in signal and image processing, economics, machine learning, and statistics. Various primal-dual methods have been rediscovered and extended, not only from algorithmic perspectives, but also from theoretical convergence guarantees. Despite of this great attempt in the algorithmic development, the corresponding supporting theory has not been well-developed, especially, the algorithms with rigorous convergence guarantees and low complexity-per-iteration.

Perhaps, applying first order methods to the dual is the most nature approach to solve constrained problems of the form (1). By means of the Lagrange duality theory, we can formulate the dual problem of (1) as a convex problem, where existing convex optimization techniques can be applied to solve it. Depending on the structure assumptions imposing on (1), the dual problem possesses useful properties that can be exploited to develop algorithms for the dual. For instance, we can use subgradient, gradient, proximal-gradient, as well as other proximal and splitting techniques to solve this problem. Then, the primal solutions of (1) can be recovered from the dual solutions [10, 20]. Among many other primal-dual splitting methods, alternating minimization algorithm (AMA) proposed by Tseng [18] becomes one of the most popular and powerful methods to solve (1) when g and h are nonsmooth and convex, and either g or h is strongly convex. Unfortunately, to the best of our knowledge, there has existed no optimization scheme to recover primal solutions of (1) in AMAs with convergence rate guarantees on both the primal objective residual and the feasibility gap.

If g and h are nonsmooth, then numerical methods for solving (1) often rely on the proximal operators of g and h . Mathematically, a proximal operator of a proper, closed, and convex function $\varphi : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}$ is defined as:

$$\text{prox}_{\varphi}(\mathbf{x}) := \underset{\mathbf{z}}{\operatorname{argmin}} \left\{ \varphi(\mathbf{z}) + (1/2)\|\mathbf{z} - \mathbf{x}\|^2 \right\}. \quad (2)$$

If prox_{φ} can be computed efficiently, i.e., by a closed form or by a polynomial time algorithm, then we say that φ has a “tractable proximity” operator. There exist many smooth and nonsmooth convex functions with tractable proximity operators as indicated in, e.g., [6, 14]. The proximal operator is in fact a special case of the resolvent in monotone inclusions [16]. Principally, the optimality condition for (1) can be cast into a monotone inclusion [1, 8]. By mean of proximity operators and gradients, splitting approaches in monotone inclusions can be applied to solve such a problem [7, 5, 8]. However, due to this generalization, the convergence guarantees and the convergence rates of these algorithms often achieve via a primal-dual gap or residual metric joined both the primal and dual variables. Such convergence guarantees do not reveal the complexity bounds of the primal sequence for (1) at intermediate iterations when we terminate the algorithm at a desired accuracy.

Our approach in this paper is briefly described as follows. First, since we work with non-strongly convex objectives g and h , we employ Nesterov’s smoothing technique via prox-functions [13] to partially smooth the dual function. Then, we apply the forward-backward splitting method to solve the smoothed dual problem, which is exactly the AMA method in [18]. Next, we introduce a new weighted averaging scheme using the Fenchel-type operators (c.f. (7)) to generate the primal

sequence simultaneously with the dual one. We then prove convergence rate guarantees for (1) in the primal variable as opposed to the dual one as in [9]. Finally, by incorporating Nesterov's acceleration step into the forward-backward splitting method, we obtain an accelerated primal-dual variant for solving (1) with a primal convergence rate guarantee. Interestingly, we can show that the primal sequence converges to an optimal solution of (1) with the $\mathcal{O}(1/k^2)$ -optimal rate provided that only g or h is strongly convex, but not the whole function f as in accelerated dual gradient methods [10], where k is the iteration counter.

Our contributions: Our specific contributions can be summarized as follows:

- a) We propose to combine Nesterov's smoothing technique, the alternating minimization idea, and the weighted-averaging strategy to develop a new primal-dual AMA algorithm for solving (1) without strong convexity assumption on g or h . We characterize the convergence rate on the absolute primal objective residual $|f(\bar{\mathbf{x}}^k) - f^*|$ and feasibility gap $\|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\|$ for the averaging primal sequence $\{\bar{\mathbf{x}}^k\}$. By an appropriate choice of the smoothness parameter, we provide the worst-case iteration-complexity of this algorithm to obtain an ϵ -primal solution.
- b) By incorporating Nesterov's accelerated step, we develop a new accelerated primal-dual AMA variant for solving (1), and characterize its worst-case iteration-complexity which is optimal in the sense of first-order black-box models [12].
- c) When either g or h is strongly convex, we recover the standard AMA algorithm as in [9], but with our averaging strategy, we obtain the $\mathcal{O}(1/k^2)$ -convergence rate on $|f(\bar{\mathbf{x}}^k) - f^*|$ and $\|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\|$ separately for the primal problem (1), not for its dual.

Let us emphasize the following points of our contributions. First, we can view the algorithms presented in this paper as the ISTA and FISTA schemes [2] applied to the smoothed dual problem of (1) instead the original dual of (1) as in [9]. The convergence rate on the dual objective residual is well-known and standard, while the convergence rates on the primal sequence are new. Second, we adapt the weights in our averaging primal sequence (c.f. (9)) to the local Lipschitz constant via a back-tracking line-search, which potentially increases the empirical performance of the algorithms. Third, the averaging primal sequence is computed via an additional sharp-operator of $h_{\mathcal{V}}$ (c.f. (7)) instead of the current primal iterate. This computation can be done efficiently (e.g., in a closed form) when $h_{\mathcal{V}}$ has a decomposable structure.

Paper organization: The rest of this paper is organized as follows. Section 2 briefly describes standard Lagrange duality framework for (1), and shows how to apply Nesterov's smoothing idea to the dual problem. The main results are presented in Sections 3 and 4, where the two new algorithms and their convergence are provided. Section 5 is devoted to investigating the strongly convex case. Concluding remarks are given in Section 6, while technical proof is moved to the appendix.

2 Primal-dual framework and smoothing technique

First, we briefly present the Lagrange duality framework for (1). Then we show how to apply Nesterov's smoothing technique to smooth the dual function of (1).

2.1 The Lagrange primal-dual framework

Let $\mathbf{x} := (\mathbf{u}, \mathbf{v})$ denote the primal variables, and $\mathcal{D} := \{\mathbf{x} \in \mathcal{U} \times \mathcal{V} : \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}\}$ denote the feasible set of (1). We define the Lagrange function of (1) corresponding

to the linear constraint $\mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}$ as $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := g(\mathbf{u}) + h(\mathbf{v}) + \langle \boldsymbol{\lambda}, \mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v} \rangle$, where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers. Then, we can define the dual function d of (1) as

$$d(\boldsymbol{\lambda}) := \min_{\mathbf{u} \in \mathcal{U}, \mathbf{v} \in \mathcal{V}} \{g(\mathbf{u}) + h(\mathbf{v}) + \langle \boldsymbol{\lambda}, \mathbf{c} - \mathbf{A}\mathbf{u} - \mathbf{B}\mathbf{v} \rangle\}. \quad (3)$$

Clearly, d can be split into three terms $d(\boldsymbol{\lambda}) = d^1(\boldsymbol{\lambda}) + d^2(\boldsymbol{\lambda}) + \langle \mathbf{c}, \boldsymbol{\lambda} \rangle$, where

$$\begin{cases} d^1(\boldsymbol{\lambda}) := \min_{\mathbf{u} \in \mathcal{U}} \{g(\mathbf{u}) - \langle \mathbf{A}^T \boldsymbol{\lambda}, \mathbf{u} \rangle\}, \\ d^2(\boldsymbol{\lambda}) := \min_{\mathbf{v} \in \mathcal{V}} \{h(\mathbf{v}) - \langle \mathbf{B}^T \boldsymbol{\lambda}, \mathbf{v} \rangle\}. \end{cases} \quad (4)$$

Using d , we can define the dual problem of (1) as

$$d^* := \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} d(\boldsymbol{\lambda}). \quad (5)$$

We say that problem (1) satisfies the Slater condition if

$$\text{ri}(\mathcal{X}) \cap \{\mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} = \mathbf{c}\} \neq \emptyset, \quad (6)$$

where $\mathcal{X} := \mathcal{U} \times \mathcal{V}$ and $\text{ri}(\mathcal{X})$ is a the relative interior of \mathcal{X} [17].

In this paper, we require the following blanket assumptions, which are standard in convex optimization.

Assumption A.1 *The functions g and h are both proper, closed, and convex (not necessarily strongly convex). The solution set \mathcal{X}^* of (1) is nonempty. The Slater condition (6) holds for (1).*

It is well-known that, under Assumption A.1, strong duality in (1) and (5) holds, i.e., we have zero duality gap which is expressed as $f^* - d^* = 0$. Moreover, for any feasible point $(\mathbf{x}, \boldsymbol{\lambda}) \in \text{dom}(f) \times \mathbb{R}^n$ and any primal-dual solution $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ with $\mathbf{x}^* := (\mathbf{u}^*, \mathbf{v}^*) \in \mathcal{X}^*$ we have: $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}) \leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = f^* = d^* \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*)$ for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\lambda} \in \mathbb{R}^n$.

Now, let us consider the components d^1 and d^2 of (4). Indeed, we can write these components as

$$\begin{aligned} d^1(\boldsymbol{\lambda}) &= -\max_{\mathbf{u} \in \mathcal{U}} \{ \langle \mathbf{A}^T \boldsymbol{\lambda}, \mathbf{u} \rangle - g(\mathbf{u}) \} = -g_{\mathcal{U}}^*(\mathbf{A}^T \boldsymbol{\lambda}), \\ d^2(\boldsymbol{\lambda}) &= -\max_{\mathbf{v} \in \mathcal{V}} \{ \langle \mathbf{B}^T \boldsymbol{\lambda}, \mathbf{v} \rangle - h(\mathbf{v}) \} = -h_{\mathcal{V}}^*(\mathbf{B}^T \boldsymbol{\lambda}), \end{aligned}$$

where $g_{\mathcal{U}}^*$ and $h_{\mathcal{V}}^*$ are the Fenchel conjugate of $g_{\mathcal{U}} := g + \delta_{\mathcal{U}}$ and $h_{\mathcal{V}} := h + \delta_{\mathcal{V}}$, respectively [17]. If we define two multivalued maps

$$\mathbf{u}^\#(\mathbf{s}) := \operatorname{argmax}_{\mathbf{u} \in \mathcal{U}} \{ \langle \mathbf{s}, \mathbf{u} \rangle - g(\mathbf{u}) \}, \text{ and } \mathbf{v}^\#(\mathbf{s}) := \operatorname{argmax}_{\mathbf{v} \in \mathcal{V}} \{ \langle \mathbf{s}, \mathbf{v} \rangle - h(\mathbf{v}) \}, \quad (7)$$

then the solution $\mathbf{u}^*(\boldsymbol{\lambda})$ of d^1 in (4) is given by $\mathbf{u}^*(\boldsymbol{\lambda}) \in \mathbf{u}^\#(\mathbf{A}^T \boldsymbol{\lambda}) \equiv \partial g_{\mathcal{U}}^*(\mathbf{A}^T \boldsymbol{\lambda})$. Similarly, the solution $\mathbf{v}^*(\boldsymbol{\lambda})$ of d^2 in (4) is given by $\mathbf{v}^*(\boldsymbol{\lambda}) \in \mathbf{v}^\#(\mathbf{B}^T \boldsymbol{\lambda}) \equiv \partial h_{\mathcal{V}}^*(\mathbf{B}^T \boldsymbol{\lambda})$. We call $\mathbf{u}^\#$ and $\mathbf{v}^\#$ the *sharp*-operator of g and h , respectively [20]. Each *oracle call* to d queries one element of the *sharp*-operators $\mathbf{u}^\#$ and $\mathbf{v}^\#$ at a given $\boldsymbol{\lambda} \in \mathbb{R}^n$.

By using the saddle point relation, we can show that $f^* \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) = f(\mathbf{x}) - \langle \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} - \mathbf{c}, \boldsymbol{\lambda}^* \rangle \leq f(\mathbf{x}) + \|\mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} - \mathbf{c}\| \|\boldsymbol{\lambda}^*\|$ for any $\mathbf{x} \in \mathcal{X}$. Hence, we have

$$-\|\boldsymbol{\lambda}^*\| \|\mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} - \mathbf{c}\| \leq f(\mathbf{x}) - f^* \leq f(\mathbf{x}) - d(\boldsymbol{\lambda}). \quad (8)$$

In this paper, we only assume that the second dual component d^2 defined by (4) satisfies the following assumption.

Assumption A.2 *The dual component d^2 defined by (4) is finite.*

This assumption holds in particular when \mathcal{V} is bounded. Moreover, $\mathbf{v}^*(\boldsymbol{\lambda})$ is well-defined for any $\boldsymbol{\lambda} \in \mathbb{R}^n$. Throughout this paper, we assume that Assumptions A.1 and A.2 holds without referring to them again.

2.2 The primal weighted averaging sequence

Given a sequence of the primal approximation $\{\tilde{\mathbf{x}}^k\}_{k \geq 0}$, where $\tilde{\mathbf{x}}^k := (\tilde{\mathbf{u}}^k, \tilde{\mathbf{v}}^k) \in \mathcal{X}$. We define the following weighted averaging sequence $\{\bar{\mathbf{x}}^k\}$ with $\bar{\mathbf{x}}^k := (\bar{\mathbf{u}}^k, \bar{\mathbf{v}}^k)$ as

$$\bar{\mathbf{u}}^k := S_k^{-1} \sum_{i=1}^k w_i \tilde{\mathbf{u}}^i, \quad \bar{\mathbf{v}}^k := S_k^{-1} \sum_{i=0}^k w_i \tilde{\mathbf{v}}^i, \quad \text{and} \quad S_k := \sum_{i=0}^k w_i, \quad (9)$$

where $\{w_i\}_{i \geq 0} \subset \mathbb{R}_{++}$ is the corresponding weights.

To avoid storing the whole sequence $\{\tilde{\mathbf{u}}^k, \tilde{\mathbf{v}}^k\}$ in our algorithms, we can compute $\{\bar{\mathbf{x}}^k\}$ recursively as follows:

$$\bar{\mathbf{u}}^k := (1 - \tau_k) \bar{\mathbf{u}}^{k-1} + \tau_k \tilde{\mathbf{u}}^k, \quad \text{and} \quad \bar{\mathbf{v}}^k := (1 - \tau_k) \bar{\mathbf{v}}^{k-1} + \tau_k \tilde{\mathbf{v}}^k, \quad \forall k \geq 1, \quad (10)$$

where $\tau_k := \frac{w_k}{S_k} \in [0, 1]$, $\bar{\mathbf{u}}^0 := \tilde{\mathbf{u}}^0$, and $\bar{\mathbf{v}}^0 := \tilde{\mathbf{v}}^0$. Clearly, for any convex function

f , we have $f(\bar{\mathbf{x}}^k) \leq S_k^{-1} \sum_{i=0}^k w_i f(\tilde{\mathbf{x}}^i)$ by the well-known Jensen inequality.

Approximate solutions: Our goal is to approximate a solution \mathbf{x}^* of (1) by \mathbf{x}_ϵ^* in the following sense:

Definition 1 Given an accuracy level $\epsilon > 0$, a point $\mathbf{x}_\epsilon^* := (\mathbf{u}_\epsilon^*, \mathbf{v}_\epsilon^*) \in \mathcal{X}$ is said to be an ϵ -solution of (1) if

$$|f(\mathbf{x}_\epsilon^*) - f^*| \leq \epsilon \quad \text{and} \quad \|\mathbf{A}\mathbf{u}_\epsilon^* + \mathbf{B}\mathbf{v}_\epsilon^* - \mathbf{c}\| \leq \epsilon. \quad (11)$$

Here, we call $|f(\mathbf{x}_\epsilon^*) - f^*|$ the [absolute] primal objective residual and $\|\mathbf{A}\mathbf{u}_\epsilon^* + \mathbf{B}\mathbf{v}_\epsilon^* - \mathbf{c}\|$ the primal feasibility gap. The condition $\mathbf{x}_\epsilon^* \in \mathcal{X}$ is in general not restrictive since, in many cases, \mathcal{X} is a simple set (e.g., a box, a simplex, or a conic cone) so that the projection onto \mathcal{X} can exactly be guaranteed.

2.3 Smoothing the dual component

As mentioned earlier, we first focus on the non-strongly convex functions g and h . In this case, we can not directly apply the standard AMA [18] to solve (1). We smooth g by using a prox-function as follows.

A continuous and strongly convex function $p_{\mathcal{U}}$ with the strong convexity parameter $\mu_p > 0$ is called a prox-function for \mathcal{U} if $\mathcal{U} \subseteq \text{dom}(p_{\mathcal{U}})$ [13]. We consider the following smoothed function d_γ^1 for d^1 :

$$d_\gamma^1(\boldsymbol{\lambda}) := \min_{\mathbf{u} \in \mathcal{U}} \{g(\mathbf{u}) - \langle \boldsymbol{\lambda}, \mathbf{A}\mathbf{u} \rangle + \gamma p_{\mathcal{U}}(\mathbf{u})\}, \quad (12)$$

where $\gamma > 0$ is a smoothness parameter.

It is well-known that d_γ^1 is concave and smooth. Moreover, as shown in [13], its gradient is given by $\nabla d_\gamma^1(\boldsymbol{\lambda}) = -\mathbf{A}\mathbf{u}_\gamma^*(\boldsymbol{\lambda})$, which is Lipschitz continuous with the Lipschitz constant $L_{d^1}^\gamma := \frac{\|\mathbf{A}\|^2}{\gamma \mu_p}$, where $\mathbf{u}_\gamma^*(\boldsymbol{\lambda})$ is the unique solution of the minimization problem in (12). In addition, we have the following estimate

$$d_\gamma^1(\boldsymbol{\lambda}) - \gamma D_{\mathcal{U}} \leq d^1(\boldsymbol{\lambda}) \leq d_\gamma^1(\boldsymbol{\lambda}), \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^n, \quad (13)$$

where $D_{\mathcal{U}}$ is the prox-diameter of \mathcal{U} , i.e.,

$$D_{\mathcal{U}} := \sup_{\mathbf{u} \in \mathcal{U}} p_{\mathcal{U}}(\mathbf{u}). \quad (14)$$

In order to develop algorithms, we require the following additional assumption.

Assumption A.3 *The quantity $D_{\mathcal{U}}$ defined by (14) is finite, i.e., $0 \leq D_{\mathcal{U}} < +\infty$.*

Clearly, if \mathcal{U} is bounded, then Assumption A.3 is automatically satisfied. Under Assumption A.3, we consider the following convex problem:

$$d_{\gamma}^* := \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\{ d_{\gamma}(\boldsymbol{\lambda}) := d_{\gamma}^1(\boldsymbol{\lambda}) + d^2(\boldsymbol{\lambda}) + \langle \mathbf{c}, \boldsymbol{\lambda} \rangle \right\}. \quad (15)$$

Using (13), we can see that d_{γ}^* converges to d^* as $\gamma \downarrow 0^+$. Hence, (15) can be considered as an approximation to the dual problem (5). We call (15) the *smoothed* dual problem of (1).

3 The non-accelerated primal-dual alternating minimization algorithm

Since d_{γ}^1 is Lipschitz gradient, we can apply the proximal-gradient method (ISTA [2]) to solve (15). This leads to the AMA scheme presented in [9, 18].

The main iteration of the alternating minimization algorithm (AMA) [18] applying to the corresponding primal problem of (15) can be written as

$$\begin{cases} \hat{\mathbf{u}}^{k+1} := \operatorname{argmin}_{\mathbf{u} \in \mathcal{U}} \{g(\mathbf{u}) - \langle \mathbf{A}^T \hat{\boldsymbol{\lambda}}^k, \mathbf{u} \rangle + \gamma p_{\mathcal{U}}(\mathbf{u})\} = \nabla g_{\gamma}^*(\mathbf{A}^T \hat{\boldsymbol{\lambda}}^k), \\ \hat{\mathbf{v}}^{k+1} := \operatorname{argmin}_{\mathbf{v} \in \mathcal{V}} \{h(\mathbf{v}) - \langle \mathbf{B}^T \hat{\boldsymbol{\lambda}}^k, \mathbf{v} \rangle + \frac{\eta_k}{2} \|\mathbf{c} - \mathbf{A} \hat{\mathbf{u}}^{k+1} - \mathbf{B} \mathbf{v}\|^2\}, \\ \boldsymbol{\lambda}^{k+1} := \hat{\boldsymbol{\lambda}}^k + \eta_k (\mathbf{c} - \mathbf{A} \hat{\mathbf{u}}^{k+1} - \mathbf{B} \hat{\mathbf{v}}^{k+1}), \end{cases} \quad (16)$$

where $\hat{\boldsymbol{\lambda}}^k \in \mathbb{R}^n$ is given, $\eta_k > 0$ is the penalty parameter, and $g_{\gamma}(\cdot) := g(\cdot) + \gamma p_{\mathcal{U}}(\cdot)$. We define the quadratic surrogate of d^1 as follows:

$$Q_{L_k}^{\gamma}(\boldsymbol{\lambda}; \hat{\boldsymbol{\lambda}}^k) := d_{\gamma}^1(\hat{\boldsymbol{\lambda}}^k) + \langle \nabla d_{\gamma}^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k \rangle - \frac{L_k}{2} \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k\|^2. \quad (17)$$

Then the following lemma provides a key estimate to prove the convergence of the algorithms in the sequel, whose proof can be found in Appendix A.

Lemma 1 *The smoothed dual component d_{γ}^1 defined by (12) is concave and smooth. It satisfies the following estimate*

$$d_{\gamma}^1(\boldsymbol{\lambda}) + \langle \nabla d_{\gamma}^1(\boldsymbol{\lambda}), \tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda} \rangle - \frac{L_{d^1}}{2} \|\tilde{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\|^2 \leq d^1(\tilde{\boldsymbol{\lambda}}), \quad \forall \boldsymbol{\lambda}, \tilde{\boldsymbol{\lambda}} \in \mathbb{R}^n, \quad (18)$$

where $L_{d^1}^{\gamma} := \frac{\|\mathbf{A}\|^2}{\gamma \mu_p}$.

Let $\boldsymbol{\lambda}^{k+1}$ be the point generated by (16) from $\hat{\boldsymbol{\lambda}}^k$ and η_k . Then, (16) is equivalent to the forward-backward splitting scheme applying to the smoothed dual problem (15), i.e.,

$$\boldsymbol{\lambda}^{k+1} := \operatorname{prox}_{(-\eta_k d^2)} \left(\hat{\boldsymbol{\lambda}}^k + \eta_k \nabla d_{\gamma}^1(\hat{\boldsymbol{\lambda}}^k) \right). \quad (19)$$

In addition, with $Q_{L_k}^\gamma$ defined by (17), if the following condition holds

$$d_\gamma^1(\lambda^{k+1}) \geq Q_{L_k}^\gamma(\lambda^{k+1}; \hat{\lambda}^k), \quad (20)$$

then, for any $\lambda \in \mathbb{R}^n$, the following estimates hold

$$\begin{aligned} d_\gamma(\lambda^{k+1}) &\geq \ell_k^\gamma(\lambda) + \frac{1}{\eta_k} \langle \lambda^{k+1} - \hat{\lambda}^k, \lambda - \hat{\lambda}^k \rangle + \left(\frac{1}{\eta_k} - \frac{L_k}{2} \right) \|\hat{\lambda}^k - \lambda^{k+1}\|^2 \\ &\geq d_\gamma(\lambda) + \frac{1}{\eta_k} \langle \lambda^{k+1} - \hat{\lambda}^k, \lambda - \hat{\lambda}^k \rangle + \left(\frac{1}{\eta_k} - \frac{L_k}{2} \right) \|\hat{\lambda}^k - \lambda^{k+1}\|^2, \end{aligned} \quad (21)$$

where $\ell_k^\gamma(\lambda) := d_\gamma^1(\hat{\lambda}^k) + \langle \nabla d_\gamma^1(\hat{\lambda}^k), \lambda - \hat{\lambda}^k \rangle + d^2(\lambda^{k+1}) + \langle \nabla d^2(\lambda^{k+1}), \lambda - \lambda^{k+1} \rangle + \langle \mathbf{c}, \lambda \rangle$, and $\nabla d^2(\lambda^{k+1}) \in \partial d^2(\lambda^{k+1})$ is a subgradient of d^2 at λ^{k+1} .

Our next step is to recover an approximate primal solution $\bar{\mathbf{x}}^k := (\bar{\mathbf{u}}^k, \bar{\mathbf{v}}^k)$ of (1) using the weighted averaging scheme (9). Combining this strategy and (16) we can present the new primal-dual AMA algorithm as in Algorithm 1 below.

Algorithm 1 (*Primal-dual alternating minimization algorithm*)

Initialization:

1. Choose $\gamma := \frac{\epsilon}{2D_U}$, and \underline{L} such that $0 < \underline{L} \leq L_{d_1}^\gamma := \frac{\|\mathbf{A}\|^2}{\gamma\mu_p}$.
2. Choose an initial point $\lambda^0 \in \mathbb{R}^n$.
3. Set $S_{-1} := 0$, $\bar{\mathbf{u}}^{-1} := 0$ and $\bar{\mathbf{v}}^{-1} := 0$.

for $k := 0$ **to** k_{\max} **do**

4. Compute $\tilde{\mathbf{u}}^k = \hat{\mathbf{u}}^{k+1} = \mathbf{u}_\gamma^*(\lambda^k)$ defined in (12).

5. Choose $\eta_k \in \left(0, \frac{1}{L_{d_1}^\gamma}\right]$ and compute

$$\hat{\mathbf{v}}^{k+1} := \arg \min_{\mathbf{v} \in \mathcal{V}} \left\{ h(\mathbf{v}) - \langle \mathbf{B}^T \lambda^k, \mathbf{v} \rangle + \frac{\eta_k}{2} \|\mathbf{c} - \mathbf{A}\tilde{\mathbf{u}}^k - \mathbf{B}\mathbf{v}\|^2 \right\}.$$

6. Update $\lambda^{k+1} := \lambda^k + \eta_k (\mathbf{c} - \mathbf{A}\hat{\mathbf{u}}^{k+1} - \mathbf{B}\hat{\mathbf{v}}^{k+1})$.
7. Compute $\tilde{\mathbf{v}}^k := \mathbf{v}^*(\lambda^{k+1}) \in \mathbf{v}^\#(\mathbf{B}^T \lambda^{k+1})$ defined in (7).
8. Update $S_k := S_{k-1} + w_k$, with $w_k := \eta_k$, and $\tau_k := \frac{w_k}{S_k}$.
9. Update $\bar{\mathbf{u}}^k := (1 - \tau_k)\bar{\mathbf{u}}^{k-1} + \tau_k \tilde{\mathbf{u}}^k$ and $\bar{\mathbf{v}}^k := (1 - \tau_k)\bar{\mathbf{v}}^{k-1} + \tau_k \tilde{\mathbf{v}}^k$.

end for

Output: The sequence $\{\bar{\mathbf{x}}^k\}$ with $\bar{\mathbf{x}}^k := (\bar{\mathbf{u}}^k, \bar{\mathbf{v}}^k)$.

In fact, we can use the Lipschitz constant $L_{d_1}^\gamma = \frac{\|\mathbf{A}\|^4}{\gamma\mu_p}$ to compute the constant step η_k as $\eta_k := \frac{1}{L_{d_1}^\gamma}$ at Step 5. However, we can adaptively choose $\eta_k = L_k^{-1}$ via a back-tracking line-search procedure in Algorithm 1 to guarantee the condition (20), and this usually performs better in practice than the constant step-size.

Algorithm 1 requires one more *sharp* operator query of \mathbf{v} at Step 7. As mentioned earlier, when $h_\mathcal{V}$ has decomposable structures, computing this sharp operator can be done efficiently (e.g., closed form or parallel/distributed manner).

The following theorem shows the bounds on the objective residual $f(\bar{\mathbf{x}}^k) - f^*$ and the feasibility gap $\|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\|$ of (1) at $\bar{\mathbf{x}}^k$.

Theorem 1 Let $\{\bar{\mathbf{x}}^k\}$ with $\bar{\mathbf{x}}^k := (\bar{\mathbf{u}}^k, \bar{\mathbf{v}}^k)$ be the sequence generated by Algorithm 1 and $L_{d^1} := \frac{\|\mathbf{A}\|^2}{\mu_p}$. Then, the following estimates hold:

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| \leq \max \left\{ \frac{L_{d^1} \|\boldsymbol{\lambda}^0\|^2}{\gamma(k+1)} + \gamma D_{\mathcal{U}}, \frac{2L_{d^1} \|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\gamma(k+1)} + \|\boldsymbol{\lambda}^*\| \sqrt{\frac{L_{d^1} D_{\mathcal{U}}}{k+1}} \right\}, \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \leq \frac{2L_{d^1} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\gamma(k+1)} + \sqrt{\frac{L_{d^1} D_{\mathcal{U}}}{k+1}}. \end{cases} \quad (22)$$

Consequently, if we choose $\gamma := \frac{\epsilon}{2D_{\mathcal{U}}}$, which is optimal, then the worst-case iteration-complexity of Algorithm 1 to achieve the ϵ -solution $\bar{\mathbf{x}}^k$ of (1) in the sense of Definition 1 is $\mathcal{O}\left(\frac{L_{d^1} D_{\mathcal{U}}}{\epsilon^2} R_0^2\right)$, where $R_0 := \max\{2, 3\|\boldsymbol{\lambda}^*\|, 2\|\boldsymbol{\lambda}^0\|, 2\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|\}$.

Proof Since $0 < \eta_i \leq \frac{1}{L_{d^1}^*}$ by Step 5 of Algorithm 1, for any $\boldsymbol{\lambda} \in \mathbb{R}^n$, it follows from (21) that

$$\begin{aligned} d_{\gamma}(\boldsymbol{\lambda}^{i+1}) &\geq \ell_i^{\gamma}(\boldsymbol{\lambda}) + \frac{1}{\eta_i} \langle \boldsymbol{\lambda}^{i+1} - \boldsymbol{\lambda}^i, \boldsymbol{\lambda} - \boldsymbol{\lambda}^i \rangle + \frac{1}{2\eta_i} \|\boldsymbol{\lambda}^{i+1} - \boldsymbol{\lambda}^i\|^2 \\ &= \ell_i^{\gamma}(\boldsymbol{\lambda}) + \frac{1}{2\eta_i} \left[\|\boldsymbol{\lambda}^{i+1} - \boldsymbol{\lambda}\|^2 - \|\boldsymbol{\lambda}^i - \boldsymbol{\lambda}\|^2 \right], \end{aligned} \quad (23)$$

where $\ell_i^{\gamma}(\boldsymbol{\lambda}) := d_{\gamma}^1(\hat{\boldsymbol{\lambda}}^i) + \langle \nabla d_{\gamma}^1(\hat{\boldsymbol{\lambda}}^i), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^i \rangle + d^2(\boldsymbol{\lambda}^{i+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{i+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{i+1} \rangle + \langle \mathbf{c}, \boldsymbol{\lambda} \rangle$ and $\nabla d^2(\boldsymbol{\lambda}^{i+1}) \in \partial d^2(\boldsymbol{\lambda}^{i+1})$ is a subgradient of d^2 at $\boldsymbol{\lambda}^{i+1}$.

Next, we consider $\ell_i^{\gamma}(\boldsymbol{\lambda})$. We first note that, for any $i = 0, \dots, k$, we have

$$\begin{aligned} d_{\gamma}^1(\boldsymbol{\lambda}^i) + \langle \nabla d_{\gamma}^1(\boldsymbol{\lambda}^i), \boldsymbol{\lambda} - \boldsymbol{\lambda}^i \rangle &= g(\hat{\mathbf{u}}^{i+1}) + \gamma p_{\mathcal{U}}(\hat{\mathbf{u}}^{i+1}) - \langle \mathbf{A}\hat{\mathbf{u}}^{i+1}, \boldsymbol{\lambda}^i \rangle - \langle \mathbf{A}\hat{\mathbf{u}}^{i+1}, \boldsymbol{\lambda} - \boldsymbol{\lambda}^i \rangle \\ &= g(\hat{\mathbf{u}}^{i+1}) - \langle \mathbf{A}\hat{\mathbf{u}}^{i+1}, \boldsymbol{\lambda} \rangle + \gamma p_{\mathcal{U}}(\hat{\mathbf{u}}^{i+1}). \end{aligned} \quad (24)$$

Second, by Step 6 of Algorithm 1, we have $\tilde{\mathbf{v}}^i \in \mathbf{v}^{\sharp}(\mathbf{B}^T \boldsymbol{\lambda}^{i+1})$, which implies

$$\begin{aligned} d^2(\boldsymbol{\lambda}^{i+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{i+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{i+1} \rangle &= h(\tilde{\mathbf{v}}^i) - \langle \mathbf{B}\tilde{\mathbf{v}}^i, \boldsymbol{\lambda}^{i+1} \rangle - \langle \mathbf{B}\tilde{\mathbf{v}}^i, \boldsymbol{\lambda} - \boldsymbol{\lambda}^{i+1} \rangle \\ &= h(\tilde{\mathbf{v}}^i) - \langle \mathbf{B}\tilde{\mathbf{v}}^i, \boldsymbol{\lambda} \rangle. \end{aligned} \quad (25)$$

Summing up (24) and (25) and using the definition of ℓ_i^{γ} , we obtain

$$\begin{aligned} \ell_i^{\gamma}(\boldsymbol{\lambda}) &= g(\tilde{\mathbf{u}}^i) + h(\tilde{\mathbf{v}}^i) - \langle \mathbf{A}\tilde{\mathbf{u}}^i + \mathbf{B}\tilde{\mathbf{v}}^i - \mathbf{c}, \hat{\boldsymbol{\lambda}}^i \rangle + \langle \mathbf{c} - \mathbf{A}\tilde{\mathbf{u}}^i - \mathbf{B}\tilde{\mathbf{v}}^i, \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^i \rangle + \gamma p_{\mathcal{U}}(\tilde{\mathbf{u}}^i) \\ &= f(\tilde{\mathbf{x}}^i) - \langle \mathbf{A}\tilde{\mathbf{u}}^i + \mathbf{B}\tilde{\mathbf{v}}^i - \mathbf{c}, \boldsymbol{\lambda} \rangle + \gamma p_{\mathcal{U}}(\tilde{\mathbf{u}}^i). \end{aligned} \quad (26)$$

By (13), we have $d_{\gamma}(\boldsymbol{\lambda}) \leq d(\boldsymbol{\lambda}) + \gamma D_{\mathcal{U}} \leq d^* + \gamma D_{\mathcal{U}} := \bar{d}_{\gamma}^*$ for any $\boldsymbol{\lambda} \in \mathbb{R}^n$. Substituting (26) into (23), subtracting to \bar{d}_{γ}^* , and summing up the result from $i = 0$ to k , we obtain

$$\begin{aligned} \sum_{i=0}^k \eta_i [\bar{d}_{\gamma}^* - d_{\gamma}(\hat{\boldsymbol{\lambda}}^{i+1})] &\leq \sum_{i=0}^k \eta_i [\bar{d}_{\gamma}^* - f(\tilde{\mathbf{x}}^i) + \langle \mathbf{A}\tilde{\mathbf{u}}^i + \mathbf{B}\tilde{\mathbf{v}}^i - \mathbf{c}, \boldsymbol{\lambda} \rangle - \gamma p_{\mathcal{U}}(\tilde{\mathbf{u}}^i)] \\ &\quad + \frac{1}{2} \left[\|\hat{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}\|^2 - \|\hat{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}\|^2 \right]. \end{aligned} \quad (27)$$

On the one hand, we note that $d(\boldsymbol{\lambda}) \leq d^* = f^* \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*) = f(\mathbf{x}) - \langle \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} - \mathbf{c}, \boldsymbol{\lambda}^* \rangle$ for any $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathcal{X}$ due to strong duality. Hence, $\langle \mathbf{A}\tilde{\mathbf{u}}^k + \mathbf{B}\tilde{\mathbf{v}}^k -$

$\mathbf{c}, \boldsymbol{\lambda}^*\rangle \leq f(\bar{\mathbf{x}}^k) - d^*$. Moreover, $\bar{d}_\gamma^* - d_\gamma(\hat{\boldsymbol{\lambda}}^{i+1}) \geq 0$. On the other hand, using the convexity of f we have $S_k f(\bar{\mathbf{x}}^k) \leq \sum_{i=0}^k w_i f(\bar{\mathbf{x}}^i)$ and $S_k \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda} \rangle = \sum_{i=0}^k w_i \langle \mathbf{A}\bar{\mathbf{u}}^i + \mathbf{B}\bar{\mathbf{v}}^i - \mathbf{c}, \boldsymbol{\lambda} \rangle$ for $w_i := \eta_i$. Combining these expressions into (27), and noting that $0 \leq p_{\mathcal{U}}(\bar{\mathbf{u}}^i) \leq D_{\mathcal{U}}$, we can derive

$$\begin{aligned} 0 &\leq \sum_{i=0}^k w_i [\bar{d}_\gamma^* - f(\bar{\mathbf{x}}^i) + \langle \mathbf{A}\bar{\mathbf{u}}^i + \mathbf{B}\bar{\mathbf{v}}^i - \mathbf{c}, \boldsymbol{\lambda} \rangle - \gamma p_{\mathcal{U}}(\bar{\mathbf{u}}^i)] + \frac{1}{2} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}\|^2 \\ &\leq S_k [d^* - f(\bar{\mathbf{x}}^k) + \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda} \rangle + \gamma D_{\mathcal{U}}] + \frac{1}{2} \|\hat{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}\|^2, \end{aligned}$$

which implies

$$\langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda}^* \rangle \leq f(\bar{\mathbf{x}}^k) - d^* \leq \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda} \rangle + \frac{1}{2S_k} \|\hat{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}\|^2 + \gamma D_{\mathcal{U}}. \quad (28)$$

Hence, we obtain

$$\langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda}^* - \boldsymbol{\lambda} \rangle - \frac{1}{2S_k} \|\hat{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}\|^2 - \gamma D_{\mathcal{U}} \leq 0, \quad (29)$$

for all $\boldsymbol{\lambda} \in \mathbb{R}^n$. Since (29) holds for all $\boldsymbol{\lambda} \in \mathbb{R}^n$, we can show that

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \left\{ \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda}^* - \boldsymbol{\lambda} \rangle - \frac{1}{2S_k} \|\hat{\boldsymbol{\lambda}}^0 - \boldsymbol{\lambda}\|^2 - \gamma D_{\mathcal{U}} \right\} \leq 0, \quad (30)$$

By optimizing the left-hand side over $\boldsymbol{\lambda} \in \mathbb{R}^n$ and using $\boldsymbol{\lambda}^0 = \hat{\boldsymbol{\lambda}}^0$, we obtain

$$S_k \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\|^2 + 2 \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c} + \mathbf{r}, \boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^* \rangle - \gamma D_{\mathcal{U}} \leq 0.$$

Using the Cauchy-Schwarz inequality, we have $\langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^* \rangle \leq \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|$. Hence, the last inequality leads to

$$\begin{aligned} \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| &\leq \frac{\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|^2 + \gamma S_k D_{\mathcal{U}}}}{S_k} \\ &\leq \frac{2\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{S_k} + \sqrt{\frac{\gamma D_{\mathcal{U}}}{S_k}}. \end{aligned} \quad (31)$$

Now, since $w_i = \eta_i \geq \frac{\gamma}{L_{d^1}}$ for $i = 0$ to k , where $L_{d^1} := \frac{\|\mathbf{A}\|^2}{\mu_p}$. Hence, $S_k \geq \frac{\gamma(k+1)}{L_{d^1}}$. Substituting this bound into (31), we obtain the second inequality of (22).

To prove the first inequality of (22), we note from (28) and $f^* = d^*$ that

$$f(\bar{\mathbf{x}}^k) - f^* \leq \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \boldsymbol{\lambda} \rangle + \frac{1}{2S_k} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}\|^2 + \gamma D_{\mathcal{U}}.$$

Taking $\boldsymbol{\lambda} = \mathbf{0}^n$ into this inequality, we get

$$f(\bar{\mathbf{x}}^k) - f^* \leq \frac{1}{2S_k} \|\boldsymbol{\lambda}^0\|^2 + \gamma D_{\mathcal{U}} \leq \frac{L_{d^1}}{\gamma(k+1)} \|\boldsymbol{\lambda}^0\|^2 + \gamma D_{\mathcal{U}}.$$

Combining this inequality, (8), and the second estimate of (22), we obtain the first estimate of (22).

Let us choose γ such that $\frac{2L_{d^1}r_0}{\gamma(k+1)} = \sqrt{\frac{L_{d^1}D_{\mathcal{U}}}{k+1}}$, where $r_0 := \max \{\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|, \|\boldsymbol{\lambda}^0\|\}$.

Then, $\gamma = \frac{2r_0\sqrt{L_{d^1}}}{\sqrt{D_{\mathcal{U}}(k+1)}}$. Substituting this expression into (22), we obtain

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| & \leq \max \left\{ \frac{2r_0 \sqrt{L_{d^1} D_{\mathcal{U}}}}{\sqrt{k+1}}, \frac{3\|\boldsymbol{\lambda}^*\| \sqrt{L_{d^1} D_{\mathcal{U}}}}{\sqrt{k+1}} \right\} \leq \epsilon \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| & \leq \frac{3\sqrt{L_{d^1} D_{\mathcal{U}}}}{\sqrt{k+1}} \leq \epsilon. \end{cases}$$

Consequently, we obtain the worst-case complexity of Algorithm 1 from the last estimates, which is $\mathcal{O}\left(\frac{L_{d^1} D_{\mathcal{U}}}{\epsilon^2} R_0^2\right)$, where $R_0 := \max\{2, 3\|\boldsymbol{\lambda}^*\|, 2\|\boldsymbol{\lambda}^0\|, 2\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|\}$. In this case, we can also show that $\gamma = \frac{\epsilon}{2D_{\mathcal{U}}}$. \square

Remark 1 If we apply a back-tracking line-search with a bi-section strategy on η_k , then we have $0 < \eta_k \leq \frac{2}{L_{d^1}^2}$ at Step 5 of Algorithm 1. In this case, the bounds in Theorem 1 still hold with $L_{d^1} = \frac{2\|\mathbf{A}\|^2}{\mu_p}$ instead of $L_{d^1} = \frac{\|\mathbf{A}\|^2}{\mu_p}$.

4 The accelerated primal-dual alternating minimization algorithm

In this section, we incorporate Nesterov's accelerated step into Algorithm 1 as done in [9], but applying to (15) to obtain a new accelerated primal-dual AMA variant. Clearly, this algorithm can be viewed as the FISTA scheme [2] applying to the smoothed dual problem (15).

Let $t_0 := 1$ and $\hat{\boldsymbol{\lambda}}^0 := \boldsymbol{\lambda}^0 \in \mathbb{R}^n$. The main step at the iteration k of the accelerated AMA method is presented as follows:

$$\begin{cases} \hat{\mathbf{u}}^{k+1} := \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} \left\{ g(\mathbf{u}) - \langle \mathbf{A}^T \hat{\boldsymbol{\lambda}}^k, \mathbf{u} \rangle + \gamma p_{\mathcal{U}}(\mathbf{u}) \right\} = \nabla g_{\gamma}^*(\mathbf{A}^T \hat{\boldsymbol{\lambda}}^k), \\ \hat{\mathbf{v}}^{k+1} := \underset{\mathbf{v} \in \mathcal{V}}{\operatorname{argmin}} \left\{ h(\mathbf{v}) - \langle \mathbf{B}^T \hat{\boldsymbol{\lambda}}^k, \mathbf{v} \rangle + \frac{\eta_k}{2} \|\mathbf{c} - \mathbf{A}\hat{\mathbf{u}}^{k+1} - \mathbf{B}\mathbf{v}\|^2 \right\}, \\ \boldsymbol{\lambda}^{k+1} := \hat{\boldsymbol{\lambda}}^k + \eta_k (\mathbf{c} - \mathbf{A}\hat{\mathbf{u}}^{k+1} - \mathbf{B}\hat{\mathbf{v}}^{k+1}), \\ t_{k+1} := \frac{1}{2} (1 + \sqrt{1 + 4t_k^2}), \\ \hat{\boldsymbol{\lambda}}^{k+1} := \boldsymbol{\lambda}^{k+1} + \frac{t_k - 1}{t_{k+1}} (\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k), \end{cases} \quad (32)$$

where, again, $g_{\gamma}(\cdot) := g(\cdot) + \gamma p_{\mathcal{U}}(\cdot)$. We now combine the accelerated AMA step (32) and the weighted averaging scheme (9) to construct a new accelerated primal-dual AMA method as presented in Algorithm 2 below.

Similar to Algorithm 1, if we know the Lipschitz constant $L_{d^1}^{\gamma}$ a priori, we can use $\eta_k := \frac{1}{L_{d^1}^{\gamma}}$. However, we can also use a backtracking line-search to adaptively choose $\eta_k := L_k^{-1}$ such that the condition (20) holds. We note that the complexity-per-iteration of Algorithm 2 essentially remains the same as in Algorithm 1.

The following theorem provides the bound on the absolute objective residual and the primal feasibility gap at the iteration $\bar{\mathbf{x}}^k$ for Algorithm 2.

Theorem 2 *Let $\{\bar{\mathbf{x}}^k\}$ be the sequence generated by Algorithm 2 and $L_{d^1} := \frac{\|\mathbf{A}\|^2}{\mu_p}$. Then, the following estimates hold:*

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| \leq \max \left\{ \frac{2L_{d^1} \|\boldsymbol{\lambda}^0\|^2}{\gamma(k+1)(k+2)} + \gamma D_{\mathcal{U}}, \frac{8L_{d^1} \|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\gamma(k+1)(k+2)} + \|\boldsymbol{\lambda}^*\| \sqrt{\frac{4L_{d^1} D_{\mathcal{U}}}{(k+1)(k+2)}} \right\}, \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \leq \frac{8L_{d^1} \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\gamma(k+1)(k+2)} + \sqrt{\frac{4L_{d^1} D_{\mathcal{U}}}{(k+1)(k+2)}}. \end{cases} \quad (33)$$

Consequently, if we choose $\gamma := \frac{\epsilon}{D_{\mathcal{U}}}$, which is optimal, then the worst-case iteration-complexity of Algorithm 2 to achieve an ϵ -solution $\bar{\mathbf{x}}^k$ of (1) in the sense of Definition 1 is $\mathcal{O}\left(\frac{\sqrt{L_{d^1} D_{\mathcal{U}}}}{\epsilon} R_0\right)$, where $R_0 := \max\{4, \frac{9}{2}\|\boldsymbol{\lambda}^0\|, \frac{9}{2}\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|, 4\|\boldsymbol{\lambda}^*\|\}$.

Algorithm 2 (*Accelerated primal-dual alternating minimization algorithm*)**Initialization:**

1. Choose $\gamma := \frac{\epsilon}{D_U}$, and \underline{L} such that $0 < \underline{L} \leq L_{d^1}^\gamma := \frac{\|\mathbf{A}\|^2}{\gamma \mu_p}$.
2. Choose an initial point $\boldsymbol{\lambda}^0 \in \mathbb{R}^n$.
3. Set $t_0 := 1$ and $\hat{\boldsymbol{\lambda}}^0 := \boldsymbol{\lambda}^0$. Set $S_{-1} := 0$, $\bar{\mathbf{u}}^{-1} := 0$ and $\bar{\mathbf{v}}^{-1} := 0$.
- for** $k := 0$ **to** k_{\max} **do**
4. Compute $\tilde{\mathbf{u}}^k = \hat{\mathbf{u}}^{k+1} = \mathbf{u}_\gamma^*(\hat{\boldsymbol{\lambda}}^k)$ defined in (15).
5. Choose $\eta_k \in \left(0, \frac{1}{L_{d^1}^\gamma}\right]$ and compute

$$\hat{\mathbf{v}}^{k+1} := \arg \min_{\mathbf{v} \in \mathcal{V}} \left\{ h(\mathbf{v}) - \langle \mathbf{B}^T \hat{\boldsymbol{\lambda}}^k, \mathbf{v} \rangle + \frac{\eta_k}{2} \|\mathbf{A} \tilde{\mathbf{u}}^k + \mathbf{B} \mathbf{v} - \mathbf{c}\|^2 \right\}.$$

6. Update $\boldsymbol{\lambda}^{k+1} := \hat{\boldsymbol{\lambda}}^k + \eta_k(\mathbf{c} - \mathbf{A} \hat{\mathbf{u}}^{k+1} - \mathbf{B} \hat{\mathbf{v}}^{k+1})$.
7. Update $t_{k+1} := 0.5(1 + (1 + 4t_k^2)^{1/2})$ and $\hat{\boldsymbol{\lambda}}^{k+1} := \boldsymbol{\lambda}^{k+1} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k)$.
8. Compute $\tilde{\mathbf{v}}^k := \mathbf{v}^*(\boldsymbol{\lambda}^{k+1}) \in \mathbf{v}^\#(\mathbf{B}^T \boldsymbol{\lambda}^{k+1})$ defined in (7).
9. Update $S_k := S_{k-1} + w_k$, with $w_k := \eta_k t_k$, and $\tau_k := \frac{w_k}{S_k}$.
10. Update $\bar{\mathbf{u}}^k := (1 - \tau_k)\bar{\mathbf{u}}^{k-1} + \tau_k \tilde{\mathbf{u}}^k$ and $\bar{\mathbf{v}}^k := (1 - \tau_k)\bar{\mathbf{v}}^{k-1} + \tau_k \tilde{\mathbf{v}}^k$.

end for

Output: The primal sequence $\{\bar{\mathbf{x}}^k\}$ with $\bar{\mathbf{x}}^k := (\bar{\mathbf{u}}^k, \bar{\mathbf{v}}^k)$.

Proof If we define $\tau_k := \frac{1}{t_k}$, then $\tau_0 = 1$, and by Step 7 of Algorithm 2, one has $\tau_{k+1}^2 = (1 - \tau_{k+1})\tau_k^2$. Moreover, if we define $\tilde{\boldsymbol{\lambda}}^k := \frac{1}{\tau_k}(\hat{\boldsymbol{\lambda}}^k - (1 - \tau_k)\boldsymbol{\lambda}^k)$, then $\tilde{\boldsymbol{\lambda}}^0 = \hat{\boldsymbol{\lambda}}^0 = \boldsymbol{\lambda}^0$. Using Step 7 of Algorithm 2, we can also derive $\tilde{\boldsymbol{\lambda}}^{k+1} = \frac{1}{\tau_{k+1}}(\hat{\boldsymbol{\lambda}}^{k+1} - (1 - \tau_{k+1})\boldsymbol{\lambda}^{k+1}) = \tilde{\boldsymbol{\lambda}}^k - \frac{1}{\tau_k}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k)$.

By (13), we have $d_\gamma(\boldsymbol{\lambda}) \leq d(\boldsymbol{\lambda}) + \gamma D_U \leq d^* + \gamma D_U := \bar{d}_\gamma^*$. Hence, $\bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}) \geq 0$ for any $\boldsymbol{\lambda} \in \mathbb{R}^n$. For $i = 0, \dots, k$, let $\ell_i^\gamma(\boldsymbol{\lambda}) := d_\gamma^1(\hat{\boldsymbol{\lambda}}^i) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^i), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^i \rangle + d^2(\boldsymbol{\lambda}^{i+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{i+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{i+1} \rangle + \langle \mathbf{c}, \boldsymbol{\lambda} \rangle$. Then, from (21) with $0 < \eta_i \leq \gamma L_{d^1}^{-1}$, and $\ell_i^\gamma(\boldsymbol{\lambda}^i) = d_\gamma^1(\hat{\boldsymbol{\lambda}}^i) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^i), \boldsymbol{\lambda}^i - \hat{\boldsymbol{\lambda}}^i \rangle + d^2(\boldsymbol{\lambda}^{i+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{i+1}), \boldsymbol{\lambda}^i - \boldsymbol{\lambda}^{i+1} \rangle + \langle \mathbf{c}, \boldsymbol{\lambda}^i \rangle \geq d_\gamma^1(\boldsymbol{\lambda}^i) + d^2(\boldsymbol{\lambda}^i) + \langle \mathbf{c}, \boldsymbol{\lambda}^i \rangle = d_\gamma(\boldsymbol{\lambda}^i)$, we have

$$\begin{aligned} \bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^{i+1}) &\leq \bar{d}_\gamma^* - \ell_i^\gamma(\boldsymbol{\lambda}) - \eta_i^{-1} \langle \boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i, \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^i \rangle - \frac{1}{2\eta_i} \|\boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i\|^2, \\ \bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^{i+1}) &\leq \bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^i) - \eta_i^{-1} \langle \boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i, \boldsymbol{\lambda}^i - \hat{\boldsymbol{\lambda}}^i \rangle - \frac{1}{2\eta_i} \|\boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i\|^2. \end{aligned} \quad (34)$$

Multiplying the first inequality of (34) by τ_i and the second one by $(1 - \tau_i)$ for $\tau_i \in (0, 1)$ and summing the results up, we obtain

$$\begin{aligned} \bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^{i+1}) &\leq (1 - \tau_i)[\bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^i)] + \tau_i[\bar{d}_\gamma^* - \ell_i^\gamma(\boldsymbol{\lambda})] \\ &\quad + \frac{1}{\eta_i} \langle \boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i, \hat{\boldsymbol{\lambda}}^i - (1 - \tau_i)\boldsymbol{\lambda}^i - \tau_i\boldsymbol{\lambda} \rangle - \frac{1}{2\eta_i} \|\boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i\|^2 \\ &= (1 - \tau_i)[\bar{d}_\gamma^* - d_\gamma(\boldsymbol{\lambda}^i)] + \tau_i[\bar{d}_\gamma^* - \ell_i^\gamma(\boldsymbol{\lambda})] \\ &\quad + \frac{\tau_i}{2\eta_i} \left[\|\tilde{\boldsymbol{\lambda}}^i - \boldsymbol{\lambda}\|^2 - \|\tilde{\boldsymbol{\lambda}}^i - \frac{1}{\tau_i}(\boldsymbol{\lambda}^{i+1} - \hat{\boldsymbol{\lambda}}^i) - \boldsymbol{\lambda}\|^2 \right], \end{aligned} \quad (35)$$

where $\tilde{\lambda}^i := \frac{1}{\tau_i}(\hat{\lambda}^i - (1 - \tau_i)\lambda^i)$. Now, let $\tilde{\lambda}^{i+1} = \tilde{\lambda}^i - \frac{1}{\tau_i}(\lambda^{i+1} - \hat{\lambda}^i)$ as stated above. Then, (35) leads to

$$\bar{d}_\gamma^* - d_\gamma(\lambda^{i+1}) \leq (1 - \tau_i) [\bar{d}_\gamma^* - d_\gamma(\lambda^i)] + \tau_i [\bar{d}_\gamma^* - \ell_i^\gamma(\lambda)] + \frac{\tau_i^2}{2\eta_i} [\|\tilde{\lambda}^i - \lambda\|^2 - \|\tilde{\lambda}^{i+1} - \lambda\|^2].$$

Now, since $\tau_i^2 = (1 - \tau_i)\tau_{i-1}^2$ and $\eta_i \leq \eta_{i-1}$, we have $\frac{\eta_i(1-\tau_i)}{\tau_i^2} \leq \frac{\eta_{i-1}}{\tau_{i-1}^2}$. Then, since $\bar{d}_\gamma^* - d_\gamma(\lambda^i) \geq 0$, the last inequality implies

$$\begin{aligned} \frac{\eta_i}{\tau_i^2} [\bar{d}_\gamma^* - d_\gamma(\lambda^{i+1})] &\leq \frac{\eta_{i-1}}{\tau_{i-1}^2} [\bar{d}_\gamma^* - d_\gamma(\lambda^i)] + \frac{\eta_i}{\tau_i} [\bar{d}_\gamma^* - \ell_i^\gamma(\lambda)] \\ &\quad + \frac{1}{2} [\|\tilde{\lambda}^i - \lambda\|^2 - \|\tilde{\lambda}^{i+1} - \lambda\|^2]. \end{aligned}$$

Summing up this inequality from $i = 0$ to k , and using the fact that $\tau_0 = 1$, we obtain

$$\begin{aligned} \frac{\eta_k}{\tau_k} [\bar{d}_\gamma^* - d_\gamma(\lambda^{k+1})] &\leq \frac{\eta_0(1-\tau_0)}{\tau_0^2} [\bar{d}_\gamma^* - d_\gamma(\lambda^k)] + \sum_{i=0}^k \frac{\eta_i}{\tau_i} [\bar{d}_\gamma^* - \ell_i^\gamma(\lambda)] \\ &\quad + \frac{1}{2} [\|\tilde{\lambda}^0 - \lambda\|^2 - \|\tilde{\lambda}^{k+1} - \lambda\|^2] \\ &\leq \sum_{i=0}^k \frac{\eta_i}{\tau_i} [\bar{d}_\gamma^* - \ell_i^\gamma(\lambda)] + \frac{1}{2} \|\tilde{\lambda}^0 - \lambda\|^2. \end{aligned} \quad (36)$$

Similar to the proof of (26), we have

$$\ell_i^\gamma(\lambda) = g(\tilde{\mathbf{u}}^i) + h(\tilde{\mathbf{v}}^i) - \langle \mathbf{A}\tilde{\mathbf{u}}^i + \mathbf{B}\tilde{\mathbf{v}}^i - \mathbf{c}, \lambda \rangle + \gamma p_{\mathcal{U}}(\tilde{\mathbf{u}}^i).$$

Next, using the convexity of g and h , and $p_{\mathcal{U}}(\tilde{\mathbf{u}}^i) \geq 0$, the last inequality implies

$$\begin{aligned} \sum_{i=0}^k \frac{\eta_i}{\tau_i} [\bar{d}_\gamma^* - \ell_i^\gamma(\lambda)] &= \sum_{i=0}^k \frac{\eta_i}{\tau_i} [\bar{d}_\gamma^* - g(\tilde{\mathbf{u}}^i) - h(\tilde{\mathbf{v}}^i) + \langle \mathbf{A}\tilde{\mathbf{u}}^i + \mathbf{B}\tilde{\mathbf{v}}^i - \mathbf{c}, \lambda \rangle - \gamma p_{\mathcal{U}}(\tilde{\mathbf{u}}^i)] \\ &\leq S_k [\bar{d}_\gamma^* - g(\bar{\mathbf{u}}^k) - h(\bar{\mathbf{v}}^k) + \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \lambda \rangle]. \end{aligned} \quad (37)$$

Substituting (37) into (36) and noting that $\bar{d}_\gamma^* \geq d_\gamma(\lambda^{k+1})$, $f(\bar{\mathbf{x}}^k) = g(\bar{\mathbf{u}}^k) + h(\bar{\mathbf{v}}^k)$ and $f^* = d^* = \bar{d}_\gamma^* - \gamma D_{\mathcal{U}}$, we have

$$f(\bar{\mathbf{x}}^k) - f^* \leq \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \lambda \rangle + \frac{1}{2S_k} \|\tilde{\lambda}^0 - \lambda\|^2 + \gamma D_{\mathcal{U}}. \quad (38)$$

Moreover, we have $f^* \leq \mathcal{L}(\mathbf{x}, \lambda^*) = f(\mathbf{x}) - \langle \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v} - \mathbf{c}, \lambda^* \rangle$ for $\mathbf{x} \in \mathcal{X}$. Substituting $\mathbf{x} := \bar{\mathbf{x}}^k$, $\mathbf{u} := \bar{\mathbf{u}}^k$ and $\mathbf{v} := \bar{\mathbf{v}}^k$ into this inequality we get

$$f^* \leq f(\bar{\mathbf{x}}^k) - \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \lambda^* \rangle. \quad (39)$$

Combining (38) and (39), we obtain

$$\langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \lambda^* - \lambda \rangle - \frac{1}{2S_k} \|\tilde{\lambda}^0 - \lambda\|^2 - \gamma D_{\mathcal{U}} \leq 0, \quad \forall \lambda \in \mathbb{R}^n. \quad (40)$$

Hence, by maximizing the left-hand side over $\lambda \in \mathbb{R}^n$, we finally get

$$\max_{\lambda \in \mathbb{R}^n} \left\{ \langle \mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}, \lambda^* - \lambda \rangle - \frac{1}{2S_k} \|\tilde{\lambda}^0 - \lambda\|^2 - \gamma D_{\mathcal{U}} \right\} \leq 0,$$

Solving the maximization problem in this inequality, we can show that

$$\|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \leq \frac{2\|\lambda^0 - \lambda^*\|}{S_k} + \sqrt{\frac{\gamma D_{\mathcal{U}}}{S_k}}. \quad (41)$$

We note that t_k updated by Step 6 satisfies: $\frac{k+1}{2} \leq t_k \leq k+1$, and $0 < \eta_k \leq \gamma L_{d^1}^{-1}$. Hence, $S_k = \sum_{i=0}^k w_i = \sum_{i=0}^k t_i \eta_i \geq \gamma \sum_{i=0}^k \frac{i+1}{2L_{d^1}} = \frac{\gamma(k+1)(k+2)}{4L_{d^1}}$. Using this estimate into (41), we get the second estimate of (33).

To prove the first estimate of (33), we note from (38) with $\lambda := \mathbf{0}^n$ that

$$f(\bar{\mathbf{x}}^k) - f^* \leq \frac{1}{2S_k} \|\lambda^0\|^2 + \gamma D_{\mathcal{U}} \leq \frac{2L_{d^1}}{\gamma(k+1)(k+2)} \|\lambda^0\|^2 + \gamma D_{\mathcal{U}}.$$

Combining this estimate, the second estimate of (33), and (8), we obtain the first estimate of (33).

Let us choose $\gamma > 0$ such that $\frac{8L_{d^1}r_0}{\gamma(k+1)(k+2)} = \sqrt{\frac{4L_{d^1}D_{\mathcal{U}}}{(k+1)(k+2)}}$, where $r_0 := \max\{\|\lambda^0\|, \|\lambda^0 - \lambda^*\|\}$. Then, $\gamma = \frac{4r_0\sqrt{L_{d^1}}}{\sqrt{D_{\mathcal{U}}(k+1)(k+2)}}$. Substituting this γ into (33), we obtain

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| & \leq \max \left\{ \frac{9r_0\sqrt{L_{d^1}D_{\mathcal{U}}}}{2\sqrt{(k+1)(k+2)}}, \frac{4\|\lambda^*\|\sqrt{L_{d^1}D_{\mathcal{U}}}}{\sqrt{(k+1)(k+2)}} \right\} \leq \epsilon \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| & \leq \frac{4\sqrt{L_{d^1}D_{\mathcal{U}}}}{\sqrt{(k+1)(k+2)}} \leq \epsilon. \end{cases}$$

Hence, the worst-case complexity of Algorithm 2 to achieve the ϵ -solution $\bar{\mathbf{x}}^k$ is $\mathcal{O}\left(\frac{\sqrt{L_{d^1}D_{\mathcal{U}}}}{\epsilon} R_0\right)$, where $R_0 := \max\{4, \frac{9}{2}\|\lambda^0\|, \frac{9}{2}\|\lambda^0 - \lambda^*\|, 4\|\lambda^*\|\}$. In this case, we also have $\gamma = \frac{\epsilon}{D_{\mathcal{U}}}$. \square

Remark 2 We note that the bounds in Theorems 1 and 2 only essentially depend on the prox-diameter $D_{\mathcal{U}}$ of \mathcal{U} , but not of \mathcal{V} . Since we can exchange g and h in the alternating step, we can choose \mathcal{U} or \mathcal{V} that has smaller prox-diameter in our algorithms to smooth its corresponding objective.

5 Application to strongly convex objectives

We assume that either g or h is strongly convex. Without loss of generality, we can assume that g is strongly convex with the convexity parameter $\mu_g > 0$ but h remains non-strongly convex, then the dual component d^1 is concave and smooth. Its gradient $\nabla d^1(\lambda) = -\mathbf{A}\mathbf{u}^*(\lambda)$ is Lipschitz continuous with the Lipschitz constant $L_{d^1} := \frac{\|\mathbf{A}\|^2}{\mu_g}$. In this case, we can modified Algorithms 1 and 2 at the following steps to capture this assumption.

- Step 1: Choose \underline{L} such that $0 < \underline{L} \leq L_{d^1} := \frac{\|\mathbf{A}\|^2}{\mu_g}$.
- Step 4: Compute $\tilde{\mathbf{u}}^k = \hat{\mathbf{u}}^{k+1} = \mathbf{u}^*(\hat{\lambda}^k) = \mathbf{u}^\#(\mathbf{A}^T \hat{\lambda}^k)$ defined by (7).
- Step 5: Choose $\eta_k \in (0, L_{d^1}^{-1}]$.

We call this modification the *strongly convex variant* of Algorithms 1 and 2, respectively. In this case, we obtain the following convergence result, which is a direct consequence of Theorems 1 and 2.

Corollary 1 *Let g be strongly convex with the convexity parameter $\mu_g > 0$. Assume that $\{\bar{\mathbf{x}}^k\}$ is the sequence generated by the strongly convex variant of Algorithm 1. Then*

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| \leq \frac{\|\mathbf{A}\|^2}{\mu_g(k+1)} \max \{ \|\boldsymbol{\lambda}^0\|^2, 2\|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| \}, \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \leq \frac{2\|\mathbf{A}\|^2 \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\mu_g(k+1)}. \end{cases} \quad (42)$$

Consequently, the worst-case iteration-complexity of this variant to achieve an ϵ -solution $\bar{\mathbf{x}}^k$ of (1) is $\mathcal{O}\left(\frac{\|\mathbf{A}\|^2 R_0}{\mu_g \epsilon}\right)$, where $R_0 := \max \{ \|\boldsymbol{\lambda}^0\|^2, 2\|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| \}$.

Alternatively, assume that $\{\bar{\mathbf{x}}^k\}$ is the sequence generated by the strongly convex variant of Algorithm 2. Then

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^*| \leq \frac{2\|\mathbf{A}\|^2}{\mu_g(k+1)(k+2)} \max \{ \|\boldsymbol{\lambda}^0\|^2, 4\|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| \}, \\ \|\mathbf{A}\bar{\mathbf{u}}^k + \mathbf{B}\bar{\mathbf{v}}^k - \mathbf{c}\| \leq \frac{8\|\mathbf{A}\|^2 \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\|}{\mu_g(k+1)(k+2)}. \end{cases} \quad (43)$$

Consequently, the worst-case iteration-complexity of this variant to achieve an ϵ -solution $\bar{\mathbf{x}}^k$ of (1) is $\mathcal{O}\left(\|\mathbf{A}\| \sqrt{\frac{R_0}{\mu_g \epsilon}}\right)$, where $R_0 := \max \{ 2\|\boldsymbol{\lambda}^0\|^2, 8\|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| \}$.

Remark 3 It is important to note that, even h is not strongly convex, our accelerated primal-dual AMA algorithm still achieves the $\mathcal{O}(1/\sqrt{\epsilon})$ -worst case iteration-complexity, which is different from existing dual accelerated schemes [3, 11, 10, 15]. In addition, if h is also strongly convex, then the sharp-operator $\mathbf{v}^\sharp(\cdot)$ of $h_{\mathcal{V}}$ is well-defined and single-valued without requiring Assumption A.2.

We note that our results present in Corollary 1 can be considered as the primal-dual variants of the AMA methods in [9], while the result presented in Theorems 1 and 2 is an extension to the non-strongly convex case.

6 Concluding remarks

We have introduce a new weighted averaging scheme, and combine the AMA idea and Nesterov's smoothing technique to develop new primal-dual AMA methods, Algorithm 1 and Algorithm 2, for solving prototype constrained convex optimization problems of the form (1) without strong convexity assumption. Then, we have incorporated Nesterov's accelerated step into Algorithm 1 to improve the worst-case iteration-complexity of the primal sequence from $\mathcal{O}(1/\epsilon^2)$ (resp., $\mathcal{O}(1/\epsilon)$ to $\mathcal{O}(1/\epsilon)$ (resp., $\mathcal{O}(1/\sqrt{\epsilon})$). Our complexity bounds are directly given for the primal objective residual and the primal feasibility gap of (1), which are new. Interestingly, the $\mathcal{O}(1/\sqrt{\epsilon})$ -complexity bound is achieved with only the strong convexity of g or h , but not both of them. We will extend this idea to other splitting schemes such as alternating direction methods of multipliers and other sets of assumptions such as the Höder continuity of the dual gradient in the forthcoming work.

A Appendix: The proof of Lemma 1

The concavity and smoothness of d_1^γ is trivial [13]. In addition, the equivalence between the AMA scheme (16) and the forward-backward splitting method was proved in, e.g., [18, 9].

Let $g_{\mathcal{U},\gamma} := g^\gamma + \delta_{\mathcal{U}}$ and $h_{\mathcal{V}} := h + \delta_{\mathcal{V}}$. We first write the optimality condition for the two convex subproblems in (16) as

$$\nabla g_{\mathcal{U},\gamma}(\hat{\mathbf{u}}^{k+1}) - \mathbf{A}^T \hat{\boldsymbol{\lambda}}^k = 0, \text{ and } \nabla h_{\mathcal{V}}(\hat{\mathbf{v}}^{k+1}) - \mathbf{B}^T \hat{\boldsymbol{\lambda}}^k - \eta_k \mathbf{B}^T (\mathbf{c} - \mathbf{A} \hat{\mathbf{u}}^{k+1} - \mathbf{B} \hat{\mathbf{v}}^{k+1}).$$

Using the third line of (16) we obtain from the last expressions that

$$\nabla g_{\mathcal{U},\gamma}(\hat{\mathbf{u}}^{k+1}) = \mathbf{A}^T \hat{\boldsymbol{\lambda}}^k, \text{ and } \nabla h_{\mathcal{V}}(\hat{\mathbf{v}}^{k+1}) = \mathbf{B}^T \boldsymbol{\lambda}^{k+1},$$

which are equivalent to

$$\hat{\mathbf{u}}^{k+1} = \nabla g_{\mathcal{U},\gamma}^*(\mathbf{A}^T \hat{\boldsymbol{\lambda}}^k), \text{ and } \hat{\mathbf{v}}^{k+1} = \nabla h_{\mathcal{V}}^*(\mathbf{B}^T \boldsymbol{\lambda}^{k+1}).$$

Multiplying these expressions by \mathbf{A} and \mathbf{B} , respectively, and adding them together, and then subtracting to \mathbf{c} , we finally obtain

$$\begin{aligned} \eta_k^{-1}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k) &= \mathbf{c} - \mathbf{A} \hat{\mathbf{u}}^{k+1} - \mathbf{B} \hat{\mathbf{v}}^{k+1} \\ &= \mathbf{c} - \mathbf{A} \nabla g_{\mathcal{U},\gamma}^*(\mathbf{A}^T \hat{\boldsymbol{\lambda}}^k) - \mathbf{B} \nabla h_{\mathcal{V}}^*(\mathbf{B}^T \boldsymbol{\lambda}^{k+1}). \end{aligned} \quad (44)$$

Now, from the definition (4) of d_γ^1 and d^2 , we have $\mathbf{A} \nabla g_{\mathcal{U},\gamma}^*(\mathbf{A}^T \hat{\boldsymbol{\lambda}}^k) = \nabla d^1(\hat{\boldsymbol{\lambda}}^k)$ and $\mathbf{B} \nabla h_{\mathcal{V}}^*(\mathbf{B}^T \boldsymbol{\lambda}^{k+1}) = \nabla d^2(\boldsymbol{\lambda}^{k+1})$. Substituting these relations into (44), we get

$$\eta_k^{-1}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k) = \mathbf{c} - \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k) - \nabla d^2(\boldsymbol{\lambda}^{k+1}). \quad (45)$$

Next, under the condition (20), we can derive

$$\begin{aligned} d_\gamma^1(\hat{\boldsymbol{\lambda}}^k) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k \rangle &= d_\gamma^1(\hat{\boldsymbol{\lambda}}^k) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k \rangle + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle \\ &= Q_{L_k}^\gamma(\boldsymbol{\lambda}^{k+1}; \hat{\boldsymbol{\lambda}}^k) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle + \frac{L_k}{2} \|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\|^2 \\ &\stackrel{(20)}{\leq} d_\gamma^1(\boldsymbol{\lambda}^{k+1}) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle + \frac{L_k}{2} \|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\|^2. \end{aligned} \quad (46)$$

Let $\ell_k^\gamma(\boldsymbol{\lambda}) := d_\gamma^1(\hat{\boldsymbol{\lambda}}^k) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k \rangle + d^2(\boldsymbol{\lambda}^{k+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{k+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle + \langle \mathbf{c}, \boldsymbol{\lambda} \rangle$. Using this expression in (46), and then combining the result with (45) and $d_\gamma(\cdot) = d_\gamma^1(\cdot) + d^2(\cdot) + \langle \mathbf{c}, \cdot \rangle$, we finally get

$$\begin{aligned} \ell_k^\gamma(\boldsymbol{\lambda}) &\leq d_\gamma(\boldsymbol{\lambda}^{k+1}) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k) + \nabla d^2(\boldsymbol{\lambda}^{k+1}) - \mathbf{c}, \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle + \frac{L_k}{2} \|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\|^2 \\ &= d_\gamma(\boldsymbol{\lambda}^{k+1}) - \langle \eta_k^{-1}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle + \frac{L_k}{2} \|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\|^2 \\ &= d_\gamma(\boldsymbol{\lambda}^{k+1}) - \langle \eta_k^{-1}(\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k \rangle - \left(\frac{1}{\eta_k} - \frac{L_k}{2} \right) \|\boldsymbol{\lambda}^{k+1} - \hat{\boldsymbol{\lambda}}^k\|^2, \end{aligned}$$

which is the first inequality of (21). The second inequality of (21) follows from the first one, $d_\gamma^1(\boldsymbol{\lambda}^k) + \langle \nabla d_\gamma^1(\hat{\boldsymbol{\lambda}}^k), \boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}^k \rangle \geq d_\gamma^1(\boldsymbol{\lambda})$ and $d^2(\boldsymbol{\lambda}^{k+1}) + \langle \nabla d^2(\boldsymbol{\lambda}^{k+1}), \boldsymbol{\lambda} - \boldsymbol{\lambda}^{k+1} \rangle \geq d^2(\boldsymbol{\lambda})$ due to the concavity of d_γ^1 and d^2 , respectively. \square

References

1. Bauschke, H., Combettes, P.: Convex analysis and monotone operators theory in Hilbert spaces. Springer-Verlag (2011)
2. Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sciences* **2**(1), 183–202 (2009)
3. Beck, A., Teboulle, M.: A fast dual proximal gradient algorithm for convex minimization and applications. *Oper. Res. Letter* **42**(1), 1–6 (2014)
4. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series). Athena Scientific (1996)
5. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision* **40**(1), 120–145 (2011)
6. Combettes, P., J.-C., P.: Signal recovery by proximal forward-backward splitting. In: *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer-Verlag (2011)
7. Combettes, P.L., Vu, B.C.: Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization* **63**(9), 1289–1318 (2014)
8. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems, vol. 1–2. Springer-Verlag (2003)
9. Goldstein, T., ODonoghue, B., Setzer, S.: Fast Alternating Direction Optimization Methods. *SIAM J. Imaging Sci.* **7**(3), 1588–1623 (2012)
10. Necoara, I., Patrascu, A.: Iteration complexity analysis of dual first order methods for convex programming. *arXiv preprint arXiv:1409.1462* (2014)
11. Necoara, I., Suykens, J.: Applications of a smoothing technique to decomposition in convex optimization. *IEEE Trans. Automatic control* **53**(11), 2674–2679 (2008)
12. Nemirovskii, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley Interscience (1983)
13. Nesterov, Y.: Smooth minimization of non-smooth functions. *Math. Program.* **103**(1), 127–152 (2005)
14. Parikh, N., Boyd, S.: Proximal algorithms. *Foundations and Trends in Optimization* **1**(3), 123–231 (2013)
15. Polyak, R.A., Costa, J., Neyshabouri, J.: Dual fast projected gradient method for quadratic programming. *Optimization Letters* **7**(4), 631–645 (2013)
16. Rockafellar, R.: Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization* **14**, 877–898 (1976)
17. Rockafellar, R.T.: Convex Analysis, *Princeton Mathematics Series*, vol. 28. Princeton University Press (1970)
18. Tseng, P., Bertsekas, D.: Relaxation methods for problems with strictly convex cost and linear constraints. *Math. Oper. Research* **16**(3), 462–481 (1991)
19. Wright, S.: Primal-Dual Interior-Point Methods. SIAM Publications, Philadelphia (1997)
20. Yurtsever, A., Tran-Dinh, Q., Cevher, V.: A universal primal-dual convex optimization framework. *Proc. of 29th Annual Conference on Neural Information Processing Systems (NIPS2015)*, Montreal, Canada, 2015.